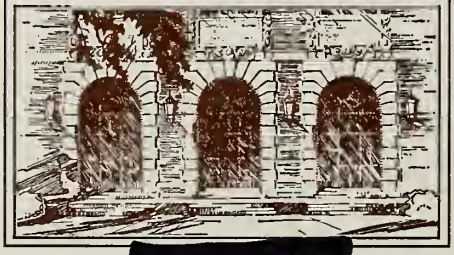

UIUCDCS-R-75-702

COO-2383-0017

ANALYSIS OF FIXED-STEPSIZE METHODS

by

Robert Skeel

February 1975

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN · URBANA, ILLINOIS

UIUCDCS-R-75-702 COO-2383-0017

ANALYSIS OF FIXED-STEPSIZE METHODS*

by

Robert Skeel

February 1975

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
URBANA, ILLINOIS 61801

* This work was supported in part by the Atomic Energy Commission under Grant US AEC AT (11-1) 2383.

PREPRINT: Submitted to SIAM Journal on Numerical Analysis.

ABSTRACT

The unified theory developed by Henrici [3] for one-step methods is extended to the more general case where the order of the system of difference equations can exceed the order of the system of ordinary differential equations. The analysis is applicable to every fixed-stepsize fixed-formula method known to the author. For many of these methods the concept of consistency is inadequate. A more appropriate concept, termed quasi-consistency, is introduced.

1. Introduction. Most numerical methods for solving initial value problems in ordinary differential equations discretize the system of differential equations and then solve the resulting system of difference equations. Henrici [3] has developed the theory for one-step methods. This paper studies the more general case where the order of the system of difference equations may exceed the order of the system of differential equations. Just as ordinary differential equations are more easily studied when written as a first order system, it is likewise profitable to study difference equations as a first order system:

$$\underline{u}_n = S\,\underline{u}_{n-1} + h\,\underline{\psi}(t_{n-1}, \underline{u}_{n-1}; h).$$

This one-step formulation is quite powerful, for it renders the analysis more transparent.

Two basic properties of methods are of practical concern: convergence, which means that the accumulated discretization error goes to zero as the step-size $h \to 0$, and stability, which ensures that the accumulated roundoff error does not grow "too fast" as $h \to 0$. Some authors (for example, Chartres and Stepleman [1]) combine the concepts of convergence and stability. This approach is cumbersome, and it does not reflect current practice in the sense that generally the roundoff error does not tend to zero like some positive power of h.

Convergent one-value (one-step) methods generate difference equations which are discrete analogs of the differential equation, and as a result the stability properties of the differential equation are preserved by the difference equation for small enough h. For multivalue methods the difference equations are not analogs of the differential equations and so the stability properties of the differential equation might not be faithfully represented. In order to ensure that this is not the case, we ought to require that a multivalue method be "as stable as" a one-value method. The notion of being "as stable as" is

defined in §3 in a quite natural way. And then in §4 it is demonstrated that being as stable as a one-value method is equivalent to requiring that the matrix S have all of its eigenvalues inside the unit circle except for an eigenvalue at 1 whose index is one and whose multiplicity is equal to the number of initial values required for the differential equation.

There are common methods, like Milne's method, which are not as stable as a one-step method but which can be expressed in a form for which they satisfy the strict root condition: all the roots of the minimal polynomial of S are inside the unit circle except possibly for a root at 1. It is determined that for methods satisfying the strict root condition convergence is equivalent to a property termed quasi-consistency. If $\underline{d}_n$, n = 0(1)N, denotes the local discretization error, then quasi-consistency means that

$$\max_{0 \leq n \leq N} | \underline{d}_n | \to 0 \quad \text{as} \quad h \to 0$$

and

$$\max_{0 \leq n \leq N} | E(\underline{d}_0 + \underline{d}_1 + \cdots + \underline{d}_n) | \to 0 \quad \text{as} \quad h \to 0$$

where $E = \lim_{m \to \infty} S^m$. On the other hand consistency usually means that

$$| \underline{d}_0 | + | \underline{d}_1 | + \cdots + | \underline{d}_N | \to 0 \quad \text{as} \quad h \to 0$$

(see Chartres and Stepleman [1], Stetter [7, pp. 5, 75]).

It is shown that the order of convergence is always equal to the order of quasi-consistency. For an often-studied class of methods which Stetter [7, p. 310] calls straight multistep methods, quasi-consistency is equivalent to consistency. This is not true for more general methods. For example, Nordsieck's [5] method is quasi-consistent, and hence convergent, of order 6 and yet only consistent of order 5. Other examples are given in §5. To have convergence of order p, it suffices to have that $\underline{d}_n = O(h^p)$ and that the essential local

truncation error $E\,\underline{d}_n = O(h^{p+1})$. In the case of linear multistep methods $E\,\underline{d}_n$ is just the local truncation error divided by $\beta_0 + \beta_1 + \cdots + \beta_q$ (cf. Gear [2, p. 118]).

In §6 the leading term in the asymptotic expansion of the global error of a method satisfying the strict root condition is determined in terms of the leading term(s) in the asymptotic expansion of the local discretization error.

2. Fixed-stepsize methods. Let s be a fixed positive integer. We consider a fixed class of systems of s first order differential equations

$$y' = f(t, y) \tag{2.1a}$$

where f is continuous in t and uniformly Lipschitz continuous in y for $0 \leq t \leq 1$, $y \in R^s$. For any initial condition

$$y(0) = \bar{y}_0 \tag{2.1b}$$

a unique solution y(t) is guaranteed. As an example, the class of systems under consideration might consist of those equations having the form

$$(y^1)' = y^2 \quad ,$$
$$(y^2)' = f(t, y^1, y^2) \quad .$$

Let q and k be fixed positive integers with $k \geq s$. The methods under consideration have three components:

(i) a mapping which to each f assigns a starting procedure

$$\underline{\sigma}(y; h) \in R^k$$

continuous for $y \in R^s$ and $0 \leq h \leq h_f$ where $0 < h_f \leq 1/q$.

(ii) a mapping, called a formula, which to each f assigns a forward step procedure

$$S\,\underline{u} + h\,\underline{\psi}(t, \underline{u}; h) \in R^k$$

where S is a $k \times k$ matrix independent of f and the increment function $\underline{\psi}$ is uniformly Lipschitz continuous in $\underline{u}$ and continuous in t and h for $0 \leq t \leq 1$, $\underline{u} \in R^k$, and $0 \leq h \leq h_f$.

(iii) a mapping[1] which to each solution y(t) of a problem assigns a correct value function

$$\underline{y}(t; h) = \Lambda\, y(t) + h\, \underline{z}(t; h) \in R^k$$

where Λ is a $k \times s$ matrix of rank s independent of y(t) and $\underline{z}(t; h)$ is bounded for $0 \leq (q-1)h \leq t \leq 1$.

---

[1] Stetter [7, p. 7] introduces a similar mapping, which he denotes by Δ_n.

Note that $y(t)$ can be recovered by means of

$$y(t) = (\Lambda^T \Lambda)^{-1} \Lambda^T \underline{y}(t; 0).$$

Given a problem and an $h \leq h_f$, a method generates a uniform grid and a discrete problem on that grid. The grid is

$$(q - 1)h = t_0 < t_1 < \ldots < t_N \leq 1$$

where $t_n = t_0 + nh$ and N is the largest integer such that $t_N \leq 1$. Since there is no danger of confusion we do not exhibit the dependence on h of such quantities as t_n and N. The discrete problem on the grid is

$$\underline{u}_0 = \underline{\sigma}(\overline{y}_0; h), \tag{2.2a}$$

$$\underline{u}_n = S\,\underline{u}_{n-1} + h\,\underline{\psi}(t_{n-1}, \underline{u}_{n-1}; h), \qquad n = 1(1)N, \tag{2.2b}$$

which can be solved to yield a numerical solution

$$\underline{U} = (\underline{u}_0, \underline{u}_1, \ldots, \underline{u}_N)^T.$$

The discrete problem (2.2) is normalized so that the unknown data vector $\underline{u}_n$ appears on the left hand side. This normalization serves to define $\underline{\sigma}$, S, and $\underline{\psi}$ uniquely.

The correct value function $\underline{y}(t; h)$ assigns a meaning to the numerical solution $\underline{U}$ by means of

$$\underline{Y} = (\underline{y}_0, \underline{y}_1, \ldots, \underline{y}_N)^T$$

where $\underline{y}_n = \underline{y}(t_n; h)$. Hence the global discretization error is defined to be $\underline{U} - \underline{Y}$.

Remark. An unfortunate aspect of methods with $k > s$ is that the data within $\underline{u}_n$ are normally inconsistent with the differential equation. For example, if $\underline{u}_n = [u_n, u_{n-1}]^T$, then there is usually no solution $y(t)$ of the differential equation such that $y(t_n) = u_n$ and $y(t_{n-1}) = u_{n-1}$.

Example 1. A straight q-step method (Stetter's [7, p. 310] terminology) determines one new value for each step:

$$\underline{u}_n = S\,\underline{u}_{n-1} + h\,\underline{e}_1\,\psi(t_{n-1}, \underline{u}_{n-1}; h)$$

where

$$\underline{u}_n = [u_n, u_{n-1}, \ldots, u_{n-q+1}]^T,$$

$$S = \begin{bmatrix} -\alpha_1 & -\alpha_2 & \cdots & -\alpha_{q-1} & -\alpha_q \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & & 1 & 0 \end{bmatrix},$$

and

$$\underline{e}_1 = [1, 0, \ldots, 0]^T.$$

Also $\underline{y}(t; h) = [y(t), y(t-h), \ldots, y(t - (q-1) h)]^T$. For the special case of a linear q-step method

$$\psi(t, \underline{u}; h) = \beta_0 \phi + \sum_{j=1}^{q} \beta_j f(t-j h, u^j)$$

where ϕ solves the equation

$$\phi = f(t, h \beta_0 \phi + \{ \sum_{j=1}^{q} (-\alpha_j u^j + h \beta_j f(t-j h, u^j))\}).$$

Example 2. Perhaps the most important class of methods which are not straight multistep methods are $P(EC)^M$ multistep methods. For example a PEC method with a second order Adams-Bashforth predictor and a third order Adams-Moulton corrector can be written

$$\begin{bmatrix} u_n \\ u_n^p \\ u_{n-1}^p \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} u_{n-1} \\ u_{n-1}^p \\ u_{n-2}^p \end{bmatrix} + h \begin{bmatrix} 5/12 & 2/3 & -1/12 \\ 0 & 3/2 & -1/2 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} f_n^p \\ f_{n-1}^p \\ f_{n-2}^p \end{bmatrix}$$

where

$$f_{n-2}^p = f(t_{n-2}, u_{n-2}^p),$$

$$f_{n-1}^p = f(t_{n-1}, u_{n-1}^p),$$

$$f_n^p = f(t_n, u_{n-1} + \frac{3}{2} h f_{n-1}^p - \frac{1}{2} h f_{n-2}^p).$$

Naturally $\underline{y}(t; h) = [y(t), y(t), y(t-h)]^T$. Alternatively it could be written

$$\begin{bmatrix} u_n \\ h\,u'_n \\ h\,u'_{n-1} \end{bmatrix} = \begin{bmatrix} 1 & 2/3 & -1/12 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} u_{n-1} \\ h\,u'_{n-1} \\ h\,u'_{n-2} \end{bmatrix} + \begin{bmatrix} 5/12 \\ 1 \\ 0 \end{bmatrix} \cdot h\,f\left(t_n,\ u_{n-1} + \frac{3}{2}h\,u'_{n-1} - \frac{1}{2}h\,u'_{n-2}\right)$$

with $\underline{y}(t; h) = [y(t), h\,y'(t), h\,y'(t-h)]^T$.

Example 3. Butcher's three-stage formula of order four (cf. Stetter [7, p. 275]) is given by

$$\begin{bmatrix} u_n \\ h\,u'_{n-1/2} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_{n-1} \\ h\,u'_{n-3/2} \end{bmatrix} + h \begin{bmatrix} 1/6 & 2/3 & 1/6 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} f_{n-1} \\ f_{n-1/2} \\ f^p_n \end{bmatrix}$$

where

$$f_{n-1} = f(t_{n-1},\ u_{n-1}),$$

$$f_{n-1/2} = f\left(t_{n-1/2},\ u_{n-1} + \frac{3}{4}h\,f_{n-1} - \frac{h}{4}u'_{n-3/2}\right),$$

$$f^p_n = f(t_n,\ u_{n-1} + 2h\,f_{n-1/2} - 2h\,f_{n-1} + h\,u'_{n-3/2})$$

and by $\underline{y}(t; h) = [y(t), h\,y'(t-h/2)]^T$.

Example 4. A modified linear multistep method equivalent to a fourth order Adams-Moulton method (Gear [2, p. 150] with $M = \infty$) is given by

$$\begin{bmatrix} u_n \\ h\,u'_n \\ \bar{u}_{n-1} \\ h\,u'_{n-1} \end{bmatrix} = \begin{bmatrix} 1/2 & 1 & 1/2 & 1/8 \\ 0 & 0 & 0 & 0 \\ 1/2 & 1/3 & 1/2 & 5/24 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} u_{n-1} \\ h\,u'_{n-1} \\ \bar{u}_{n-2} \\ h\,u'_{n-2} \end{bmatrix} + h \begin{bmatrix} 3/8 \\ 1 \\ -1/24 \\ 0 \end{bmatrix} \psi(t_{n-1},\ \underline{u}_{n-1};\ h)$$

where $\underline{y}(t;h) = [y(t), h\, y'(t), y(t-h), h\, y'(t-h)]^T$ and $\psi(t, \underline{u}; h) = \psi$ the solution of

$$\psi = f(t, \tfrac{1}{2} u^1 + u^2 + \tfrac{1}{2} u^3 + \tfrac{1}{8} u^4 + \tfrac{3}{8} h\, \psi).$$

3. Convergence and stability. Let $|\cdot|$ be a norm for R^k, and define

$$\| \underline{U} - \underline{Y} \|_\infty = \max_{0 \le n \le N} | \underline{u}_n - \underline{y}_n |.$$

DEFINITION 3.1. A method is said to be *convergent* for the problem (2.1) if

$$\| \underline{U} - \underline{Y} \|_\infty \to 0 \text{ as } h \to 0.$$

If the method is convergent for all infinitely smooth problems (2.1), then it is simply said to be convergent.

In practice $\underline{U}$ cannot be computed exactly; we must take into account the local computing error which arises from doing operations in finite precision. Hence we consider the perturbed numerical solution $\hat{\underline{U}} = (\hat{\underline{u}}_0, \hat{\underline{u}}_1, \ldots, \hat{\underline{u}}_N)^T$ defined by

$$\hat{\underline{u}}_0 = \underline{\sigma}(\overline{y}_0; h) + \underline{r}_0,$$

$$\hat{\underline{u}}_n = S\, \hat{\underline{u}}_{n-1} + h\, \underline{\psi}(t_{n-1}, \hat{\underline{u}}_{n-1}; h) + \underline{r}_n, \quad n = 1(1)N,$$

for any perturbation $\underline{R} = (\underline{r}_0, \underline{r}_1, \ldots, \underline{r}_N)^T$. The dependence of $\hat{\underline{U}}$ on $\underline{R}$ is not made explicit in our notation.

Various definitions of stability as an absolute concept have been proposed but most of them seem somewhat arbitrary. (An exception is the use of Spijker's [6] norm to define stability for straight multistep methods.) It is more natural to define stability as a relative concept.

DEFINITION 3.2. Method 2 is said to be *at least as stable* as method 1 if for each problem there is a fixed number K such that

$$\| \hat{\underline{U}}^{(2)} - \underline{U}^{(2)} \|_\infty \le K \| \hat{\underline{U}}^{(1)} - \underline{U}^{(1)} \|_\infty$$

for any $h \le h_f$ and any perturbation $\underline{R} \in R^{(N+1)k}$. If, in addition, method 2 is at least as stable as method 1, then the methods are said to be *equally stable*; otherwise method 1 is said to be *more stable* than method 2.

Since good stability properties are important for practical algorithms, it is desirable that a method satisfy the most stringent stability requirement that is compatible with convergence.

DEFINITION 3.3. A convergent method is *optimally stable* if there is no convergent method which is more stable.

Note. In general it is not possible to compare the stability of two arbitrary methods, and hence two optimally stable methods are not necessarily equally stable.

Remark. Spijker's [6] definition of optimal stability differs from ours in that it requires an optimally stable method to be at least as stable as every other method in the class of methods under consideration.

It is shown in §4 that the optimally stable methods are those convergent methods that satisfy the very strict root condition, which is defined below.

DEFINITION 3.4. A method is said to satisfy the *root condition* (RC) if the roots of the minimal polynomial of S are either inside the unit circle or on the unit circle and simple. A method satisfies the *strict root condition* (SRC) if the roots of the minimal polynomial are inside the unit circle except possibly for a simple root at 1. A method satisfies the *very strict root condition* (VSRC) if it satisfies the SRC and if S has at most s eigenvalues equal to 1.

It is shown by Theorem 3.9 that the matrix S must have at least s eigenvalues equal to 1 in order for the method to be convergent; however any additional eigenvalues on the unit circle impart a degradation to the stability of the method.

There is no increase in the complexity of the theory if instead of

assuming the VSRC we merely require that the strict root condition be satisfied. Furthermore, it is then possible to treat marginally useful methods like Milne's method. Hence the theory is developed for methods satisfying the SRC.

The following lemma gives the best possible qualitative bound on the difference $\hat{\underline{U}} - \underline{U}$ in terms of $\underline{R}$ for methods satisfying the RC. First we introduce some notation. For any $k \times k$ matrix T let [T] denote the $(N+1) \times (N+1)k$ matrix

$$\begin{bmatrix} I & & & \\ -T & I & & \\ & \ddots & \ddots & 0 \\ 0 & & \ddots & I \\ & & & -T \end{bmatrix}.$$

Then the quantity

$$|| \, [T]^{-1} \, \underline{R} \, ||_\infty = \max_{0 \le n \le N} \left| \sum_{j=0}^{n} T^{n-j} \, \underline{r}_j \right|$$

is a norm for $R^{(N+1)k}$.

LEMMA 3.5. *Given a method satisfying the RC and a problem* (2.1), *there exist positive constants* c *and* C *such that for any* $h \le h_f$ *and any* $\underline{R} \in R^{(N+1)k}$

$$c \, || \, [S]^{-1} \, \underline{R} \, ||_\infty \le || \, \hat{\underline{U}} - \underline{U} \, ||_\infty \le C \, || \, [S]^{-1} \, \underline{R} \, ||_\infty .$$

Proof. Set $\underline{\delta}_n = \hat{\underline{u}}_n - \underline{u}_n$. Then

$$\underline{\delta}_n = S \, \underline{\delta}_{n-1} + h \, \tilde{\underline{\delta}}_{n-1} + \underline{r}_n \tag{3.1}$$

where

$$\tilde{\underline{\delta}}_{n-1} = \underline{\psi}(t_{n-1}, \hat{\underline{u}}_{n-1}; h) - \underline{\psi}(t_{n-1}, \underline{u}_{n-1}; h)$$

By assumption there exists a constant L such that

$$| \, \tilde{\underline{\delta}}_{n-1} \, | \le L \, | \, \underline{\delta}_{n-1} \, |.$$

Solving the difference equation (3.1) gives

$$\underline{\delta}_n = \sum_{j=1}^{n} S^{n-j} h \tilde{\underline{\delta}}_{j-1} + \sum_{j=0}^{n} S^{n-j} \underline{r}_j. \tag{3.2}$$

By the root condition there exists a constant B such that

$$| S^n |_L \leq B. \tag{3.3}$$

Thus (3.2) becomes

$$| \underline{\delta}_n | \leq h B \sum_{j=0}^{n-1} | \underline{\delta}_j | + || [S]^{-1} \underline{R} ||_\infty.$$

By induction on n it follows that

$$| \underline{\delta}_n | \leq (1 + h B)^n || [S]^{-1} \underline{R} || \leq e^B || [S]^{-1} \underline{R} ||,$$

which proves the second inequality in the lemma. From (3.2) and (3.3)

$$| \sum_{j=0}^{n} S^{n-j} \underline{r}_j | \leq (1 + n h B) \max_{0 \leq n \leq N} | \underline{\delta}_n | \leq (1 + B) || \hat{\underline{U}} - \underline{U} ||_\infty,$$

and thus $|| S^{-1} \underline{R} ||_\infty \leq (1 + B) || \hat{\underline{U}} - \underline{U} ||_\infty$. Q.E.D.

When a norm, such as $|| [S]^{-1} \cdot ||_\infty$, permits an upper bound on $\hat{\underline{U}} - \underline{U}$, then the method is said to be *stable with respect to* that norm, and if the norm also permits a lower bound on $\hat{\underline{U}} - \underline{U}$, then the norm is said to be a *minimal stability functional* for the method. These ideas are discussed in more detail by Spijker [6] and Stetter [7, pp. 81-84].

Of considerable importance to the theory is the component of S (cf. Lancaster [4, p. 162]) corresponding to the eigenvalue 1, which we denote by E. Assuming S satisfies the RC, we have that

$$E = \overline{\rho}(1)^{-1} \overline{\rho}(S)$$

where $\rho(\xi) = (\xi - 1)\overline{\rho}(\xi)$ is the minimal polynomial of S. If S does not have an eigenvalue at 1, then we define E = 0. It is easily verified that E is idempotent and that S E = E = E S whence

$$S^n = E + (S - E)^n, \qquad n = 1, 2, \ldots .$$

If S satisfies the SRC, the eigenvalues of S - E are inside the unit circle.

The following theorem is for methods satisfying the SRC, and it gives bounds on the difference $\hat{\underline{U}} - \underline{U}$ in terms of the more useful norm

$$||[E]^{-1}\,\underline{R}\,||_\infty = \max_{0 \leq n \leq N} \left| E \sum_{j=0}^{n-1} \underline{r}_j + \underline{r}_n \right|.$$

THEOREM 3.6. *Given a method satisfying the SRC and a problem* (2.1), *there exist positive constants* c' *and* C' *such that for any* $h \leq h_f$ *and any* $\underline{R} \in R^{(N+1)k}$

$$c' \,||\,[E]^{-1}\,\underline{R}\,||_\infty \leq ||\,\hat{\underline{U}} - \underline{U}\,||_\infty \leq C' \,||\,[E]^{-1}\,\underline{R}\,||_\infty.$$

Proof. The theorem follows from Lemma 3.5 if it can be shown that $||\,[E]^{-1}\,\underline{R}\,||_\infty$ can be bounded by some constant times $||\,[S]^{-1}\,\underline{R}\,||_\infty$ and vice versa. First of all

$$\begin{aligned}
||\,[E]^{-1}\,\underline{R}\,||_\infty &= ||\,[E]^{-1}\,[S]\,[S]^{-1}\,\underline{R}\,||_\infty \\
&= ||\,[S - E]\,[S]^{-1}\,\underline{R}\,||_\infty \\
&= \max_{0 \leq n \leq N} \left| \sum_{j=0}^{N} ([S - E])_{nj}\,([S]^{-1}\,\underline{R})_j \right| \\
&\leq \max_{0 \leq n \leq N} \sum_{j=0}^{N} |\,([S - E])_{nj}\,|\;||\,[S]^{-1}\,\underline{R}\,||_\infty \\
&= (1 + |\,S - E\,|)\;||\,[S]^{-1}\,\underline{R}\,||_\infty,
\end{aligned}$$

and secondly

$$\begin{aligned}
||\,[S]^{-1}\,\underline{R}\,||_\infty &= ||\,[S]^{-1}\,[E]\,[E]^{-1}\,\underline{R}\,||_\infty \\
&= ||\,[S - E]^{-1}\,[E]^{-1}\,\underline{R}\,||_\infty \\
&= \max_{0 \leq n \leq N} \left| \sum_{j=0}^{N} ([S - E]^{-1})_{nj}\,([E]^{-1}\,\underline{R})_j \right| \\
&\leq \max_{0 \leq n \leq N} \sum_{j=0}^{N} |\,([S - E]^{-1})_{nj}\,|\;||\,[E]^{-1}\,\underline{R}\,||_\infty \\
&= (1 + |\,S - E\,| + \ldots + |\,(S - E)^N\,|)\;||\,[E]^{-1}\,\underline{R}\,||_\infty.
\end{aligned}$$

The SRC ensures that $1 + |\,S - E\,| + \ldots + |\,(S - E)^N\,|$ is bounded.

Q.E.D.

The local discretization error $\underline{D} = (\underline{d}_0, \underline{d}_1, \ldots, \underline{d}_N)^T$ is defined by

$$\underline{d}_0 = \underline{\sigma}(\bar{y}_0; h) - \underline{Y}_0 \quad ,$$

$$\underline{d}_n = S\,\underline{Y}_{n-1} + h\,\underline{\psi}(t_{n-1}, \underline{Y}_{n-1}; h) - \underline{Y}_n \quad , \quad n = 1(1)N.$$

DEFINITION 3.7. A method satisfying the RC is said to be *quasi-consistent* with the problem (2.1) if

$$\max_{0 \le n \le N} | \underline{d}_n | \to 0 \quad \text{as} \quad h \to 0$$

and

$$\max_{0 \le n \le N} | E(\underline{d}_0 + \underline{d}_1 + \ldots + \underline{d}_n) | \to 0 \quad \text{as} \quad h \to 0.$$

A method is *consistent* with (2.1) if

$$| \underline{d}_0 | + | \underline{d}_1 | + \ldots + | \underline{d}_N | \to 0 \quad \text{as} \quad h \to 0.$$

Clearly consistency implies quasi-consistency. For a method to be quasi-consistent with a problem it is sufficient that

$$\text{(i)} \quad \underline{d}_n = o(1) \quad , \quad n = 0(1)N,$$

$$\text{(ii)} \quad E\,\underline{d}_n = o(h) \quad , \quad n = 1(1)N$$

where we employ the convention that whenever the little-oh or big-oh notation is used, the implied limit or bound is uniform w.r.t.n. The quantity $E\,\underline{d}_n$ is the *essential* local discretization error.

THEOREM 3.8. *A method satisfying the SRC is convergent for* (2.1) *if and only if it is quasi-consistent with* (2.1).

Proof. When $\underline{R} = -\underline{D}$, then $\hat{\underline{U}} = \underline{Y}$; i.e., the true solution is a perturbed numerical solution. (This unified treatment of discretization error and roundoff error is due to Chartres and Stepleman [1].) Thus from Theorem 3.6 convergence is equivalent to

$$\max_{0 \le n \le N} | E(\underline{d}_0 + \underline{d}_1 + \ldots + \underline{d}_{n-1}) + \underline{d}_n | \to 0 \quad \text{as} \quad h \to 0,$$

which is equivalent to quasi-consistency. Q.E.D.

We extend a theorem of Henrici [3, p. 71] stating necessary and sufficient conditions for convergence.

THEOREM 3.9. *A method satisfying the SRC is quasi-consistent with the differential equation* (2.1a) *if and only if*

$$\text{(i)} \quad \underline{\sigma}(y;\, 0) \equiv \Lambda y,$$

$$\text{(ii)} \quad S\,\Lambda = \Lambda,$$

$$\text{(iii)} \quad E\,\underline{\psi}(t,\, \Lambda y;\, 0) \equiv E\,\Lambda\, f(t,\, y).$$

Proof. Using the fact that in the region $0 \leq t \leq 1$, $0 \leq h \leq h_f$ continuity is equivalent to uniform continuity, we have

$$\underline{d}_0 = \underline{\sigma}(\overline{y}_0;\, 0) - \Lambda\, \overline{y}_0 + o(1), \tag{3.4}$$

$$\begin{aligned} \underline{d}_n &= S\,\Lambda\, y_n - \Lambda\, y_n + O(h) \\ &= (S\,\Lambda - \Lambda) y_n + O(h), \qquad n = 1(1)N, \end{aligned} \tag{3.5}$$

and

$$\begin{aligned} E\,\underline{d}_n &= E\,\{\underline{y}_{n-1} + h\,\underline{\psi}(t_{n-1},\, \underline{y}_{n-1};\, h) - \underline{y}_n\} \\ &= E\,\{\underline{y}_{n-1} - \underline{y}_n + h\,\underline{\psi}(t_{n-1},\, \Lambda\, y_{n-1};\, 0)\} + o(h), \qquad n = 1(1)N. \end{aligned}$$

Therefore

$$\begin{aligned} \sum_{j=1}^{n} E\,\underline{d}_j &= E\,\{\underline{y}_0 - \underline{y}_n + \sum_{j=0}^{n-1} h\;\underline{\psi}(t_j,\, \Lambda\, y_j;\, 0)\} + o(1) \\ &= E\,\{\Lambda(y_0 - y_n) + \int_{t_0}^{t_n} \underline{\psi}(\tau,\, \Lambda\, y(\tau);\, 0)\, d\tau\} + o(1) \\ &= E\,\int_0^{t_n} \{-\Lambda\, f(\tau,\, y(\tau)) + \underline{\psi}(\tau,\, \Lambda y(\tau);\, 0)\} d\tau + o(1). \end{aligned} \tag{3.6}$$

Assume (i), (ii), and (iii) hold. Then it clearly follows that

$$\underline{d}_0 = o(1) \quad ,$$

$$\underline{d}_n = O(h) \quad , \quad n = 1(1)N \quad ,$$

$$\sum_{j=1}^{n} E\,\underline{d}_j = o(1) \quad , \quad n = 1(1)N,$$

which are sufficient for quasi-consistency. Now assume the method is

quasi-consistent with (2.1a). Then $\underline{d}_n = o(1)$, $n = 0(1)N$, and so from (3.4) and (3.5) we have $\underline{\sigma}(\bar{y}_0; 0) - \Lambda \bar{y}_0 = 0$ and $(S \Lambda - \Lambda)y_1 = 0$. Parts (i) and (ii) immediately follow because $\bar{y}_0$ is arbitrary and $y(t)$ is continuous. By convention equation (3.6) means that there is some function $\eta(h)$ tending to 0 as $h \to 0$ such that

$$\max_{1 \leq n \leq N} \left| E \int_0^{t_n} \{-\Lambda f(\tau, y(\tau)) + \underline{\psi}(\tau, \Lambda y(\tau); 0)\} d\tau \right| \leq \eta(h).$$

Hence for fixed $0 \leq t \leq 1$

$$\left| E \int_0^{t} \{-\Lambda f(\tau, y(\tau)) + \underline{\psi}(\tau, \Lambda y(\tau); 0)\} d\tau \right| \leq \eta\left(\frac{t}{n - q + 1}\right),$$

$n = q, q + 1, \ldots,$ and so the left-hand side must be zero. Since t is arbitrary and the integrand is continuous,

$$E \{-\Lambda f(t, y(t)) + \underline{\psi}(t, \Lambda y(t); 0)\} \equiv 0$$

By varying y_0, $y(t)$ can be made to assume any value and so (iii) follows. Q.E.D.

For a convergent method satisfying the VSRC part (ii) of Theorem 3.9 implies the existence of a unique $k \times s$ matrix M such that

$$E = \Lambda M^T.$$

The rows of M are linearly independent left eigenvalues of S corresponding to the eigenvalue 1. Part (iii) of Theorem 3.9 thus reduces to

$$M^T \underline{\psi}(t, \Lambda y; 0) = f(t, y). \tag{3.7}$$

Remark. Conditions (i)-(iii) of Theorem 3.9 are equivalent to convergence for methods which merely satisfy the RC if we strengthen the smoothness conditions on $\underline{z}(t; h)$ as follows:

(i) $\left| \underline{z}(t; h) - \underline{z}(t; 0) \right| \leq \eta(h)$ where $\eta(h) \to 0$ as $h \to 0$,

(ii) $\underline{z}(t; 0)$ is Riemann integrable (cf. Chartres and Stepleman [1, p. 484]).

We can use the fact that

$$\left| \sum_{j=0}^{n} (S - E)^j \right|$$

is bounded independently of n to show that for each problem (2.1) there exist positive constants c" and C" such that for any $h \leq h_f$ and $\underline{R} \in R^{(N+1)k}$

$$c" \; || \; [E]^{-1} \underline{R} \; ||_\infty \leq || \; \hat{\underline{U}} - \underline{U} \; ||_\infty \leq C" \; || \; \underline{R} \; ||$$

where

$$|| \; \underline{R} \; || = \max_{0 \leq n \leq N} \; | \; \sum_{j=0}^{n} E \; \underline{r}_j \; | + | \; \underline{r}_0 \; | + \sum_{n=1}^{N} | \; \underline{r}_n - \underline{r}_{n-1} \; |.$$

Just as before conditions (i) - (iii) of Theorem 3.9 are necessary for convergence. On the other hand, the additional restrictions on $\underline{z}(t; h)$ make it possible to show that

$$\sum_{n=1}^{N} | \; \underline{d}_n - \underline{d}_{n-1} \; | = o(1),$$

and hence that conditions (i) - (iii) are sufficient for convergence.

Example 1. Let us examine the quasi-consistency conditions for a straight multistep method satisfying the SRC. We have

$$\Lambda = \underline{e} = [1, 1, \ldots, 1]^T$$

and so part (ii) of Theorem 3.9 requires that

$$-\alpha_1 - \alpha_2 - \cdots - \alpha_q = 1. \tag{3.8}$$

Since the minimal polynomial of S is $\xi^q + \alpha_1 \xi^{q-1} + \cdots + \alpha_q$, the SRC is equivalent to the VSRC, and hence $E = \underline{e} \; \underline{m}^T$ where $\underline{m}^T$ is the left eigenvalue of S normalized so that $\underline{m}^T \underline{e} = 1$. It is easily verified that

$$\underline{m}^T = \overline{\rho}(1)^{-1} [1, \; 1 + \alpha_1, \; \ldots, \; 1 + \alpha_1 + \ldots + \alpha_{q-1}]$$

where

$$\overline{\rho}(\xi) = \xi^q + (1 + \alpha_1) \xi^{q-1} + \cdots + (1 + \alpha_1 + \cdots + \alpha_{q-1}).$$

Because $\underline{\psi}(t, \underline{u}; h) = \underline{e}_1 \; \psi(t, \underline{u}; h)$, we have from (3.7) that

$$\psi(t, \Lambda y; 0) = \overline{\rho}(1) \; f(t, y). \tag{3.9}$$

Together with part (i) of Theorem 3.9, (3.8) and (3.9) constitute the quasi-consistency conditions for straight multistep methods. In this case

quasi-consistency is equivalent to consistency because the local discretization error $\underline{e}_{-1}\, d_n$ and the essential local discretization error $\overline{\rho}(1)^{-1}\, \underline{e}\, d_n$ are both multiples of the same quantity d_n.

Example 2. The SRC does not exclude those weakly stable linear multistep methods based on numerical quadrature. For example, two steps of Milne's method,

$$u_n = u_{n-2} + \frac{h}{2}\left(\frac{1}{3} f_n + \frac{4}{3} f_{n-1} + \frac{1}{3} f_{n-2}\right),$$

with stepsize h/2 are equivalent to one step of the method

$$\begin{bmatrix} u_n \\ u_{n-1/2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_{n-1} \\ u_{n-3/2} \end{bmatrix} + h \begin{bmatrix} \frac{1}{6} f_n + \frac{2}{3} f_{n-1/2} + \frac{1}{6} f_{n-1} \\ \frac{1}{6} f_{n-1/2} + \frac{2}{3} f_{n-1} + \frac{1}{6} f_{n-3/2} \end{bmatrix},$$

which does satisfy the SRC though not the VSRC. The above pair of difference equations is the discretization of a system of two identical differential equations, which is not surprising because the derivation of Milne's method involves two numerical integrations over each subinterval.

Example 3. The VSRC does not exclude linear multistep methods for the special second order differential equation $y'' = f(t, y)$ as long as the difference equation is written in a form that does not have bad roundoff properties. For example the summed form of Störmer's method,

$$u_n = u_{n-1} + h\, F_{n-1},$$

$$F_n = F_{n-1} + h\, f(t_n, u_n),$$

satisfies the VSRC.

4. Optimally stable methods. We wish to characterize those methods that satisfy the most stringent definition of stability which is compatible with convergence. Spijker [6] has shown that among convergent straight multistep methods which satisfy the RC there are some which are more stable than the others and that these optimally stable methods are those which satisfy the strict root condition (which for straight multistep methods is equivalent to the very strict root condition). Theorem 4.2 extends Spijker's result by considering a much more general class of methods. In the proof of the theorem every optimally stable method is shown to be equally stable with a form of Euler's method.

LEMMA 4.1. *If a method satisfies the VSRC and if a second method is at least as stable as the first method, then the second method must satisfy the VSRC and the* E *matrices of the two methods must satisfy* $E_2 E_1 = E_2$.

Proof. Call the first method Method 1 and the second method Method 2. By Definition 3.2 and Theorem 3.6 there is a constant K independent of N such that

$$|| \hat{\underline{U}}^{(2)} - \underline{U}^{(2)} ||_\infty \leq K || [E_1]^{-1} \underline{R} ||_\infty .$$

Let m be an arbitrary fixed nonnegative integer, and let $\underline{R} = (\underline{r}_0, \underline{r}_1, \ldots, \underline{r}_m, \underline{0}, \ldots, \underline{0})^T$. Then

$$| \hat{\underline{u}}_m^{(2)} - \underline{u}_m^{(2)} | \leq K \max_{0 \leq n \leq m} | E_1(\underline{r}_0 + \ldots + \underline{r}_{n-1}) + \underline{r}_n |$$

for $0 < h < \min(h_f, 1/(m + q - 1))$. Since the left-hand side is continuous in h, we can let $h \to 0$ to get

$$(4.1) \quad | \sum_{n=0}^{m} S_2^{m-n} \underline{r}_n | \leq K \max_{0 \leq n \leq m} | E_1(\underline{r}_0, + \ldots + \underline{r}_{n-1}) + \underline{r}_n | .$$

This inequality holds for every nonnegative integer m. Choosing $\underline{r}_n = 0$, n = 1(1)m, shows that S_2 must be power-bounded, which implies the RC

is satisfied. Suppose, however, Method 2 does not satisfy the SRC. Then there must be an eigenvalue $\xi \neq 1$ of S_2 of modulus 1 and a corresponding eigenvector $\underline{v}$. Let

$$\underline{r}_n = \xi^{n-m} \underline{v} \quad , \quad n = 1(1)m.$$

Then from (4.1) we get

$$(m+1)|\underline{v}| \leq K \max_{0 \leq n \leq m} \left| \frac{\xi^{n-m} - \xi^{-m}}{\xi - 1} E_1 \underline{v} + \xi^{n-m} \underline{v} \right|$$

$$\leq K\left(\frac{2|E_1 \underline{v}|}{|\xi - 1|} + |\underline{v}| \right).$$

Since this inequality cannot hold for every m, we conclude that the SRC must be satisfied. Therefore Theorem 3.6 may be applied to Method 2, and so there is a constant $c > 0$ such that

$$(4.2) \quad c \, || [E_2]^{-1} \underline{R} ||_\infty \leq || \hat{\underline{U}}^{(2)} - \underline{U}^{(2)} ||_\infty \leq K || [E_1]^{-1} \underline{R} ||_\infty.$$

Choose $\underline{r}_n = (E_1 - I) \underline{v}$, $n = 0(1)N$, where $\underline{v}$ is arbitrary. Then (4.2) becomes

$$c \, | N E_2 (E_1 - I) \underline{v} + (E_1 - I) \underline{v} | \leq K | (E_1 - I) \underline{v} |.$$

Because both c and K are independent of N, $E_2(E_1 - I) \underline{v} = \underline{0}$, and because $\underline{v}$ is arbitrary,

$$E_2 E_1 = E_2.$$

Therefore

$$\text{rank } E_2 = \text{rank } E_2 E_1 \leq \text{rank } E_1 \leq s$$

showing that the VSRC is satisfied. Q.E.D.

THEOREM 4.2. *A convergent method is optimally stable if and only if it satisfies the VSRC.*

Proof. Let a convergent method be given. The global discretization errors are

$$\underline{\varepsilon}_0 = \underline{\sigma}(\bar{y}_0; 0) - A \bar{y}_0 + o(1),$$

$$\underline{\varepsilon}_1 = S \underline{\sigma}(\bar{y}_0; 0) - A \bar{y}_0 + o(1).$$

Convergence implies that

$$\underline{\sigma}(y;0) = \Lambda\, y,$$

$$S\,\underline{\sigma}(y;\,0) = \Lambda\, y$$

for all y; hence $S\,\Lambda = \Lambda$. This means that the columns of Λ are linearly independent right eigenvectors of S corresponding to the eigenvalue 1. Therefore there exists a $k \times s$ matrix Λ_0 of rank s such that

$$\Lambda_0^T S = \Lambda_0^T.$$

Define

$$\begin{aligned} \underline{\sigma}_0(y;\,h) &= \Lambda_0\, y, \\ S_0 &= \Lambda_0(\Lambda_0^T\,\Lambda_0)^{-1}\,\Lambda_0^T, \\ \underline{\psi}_0(t,\,\underline{u};\,h) &= \Lambda_0\, f(t,\,S_0\,\underline{u}), \\ \underline{y}_0(t;\,h) &= \Lambda_0\, y(t). \end{aligned} \tag{4.3}$$

The method (4.3) satisfies the VSRC and is convergent by Theorem 3.9. Let us show that (4.3) is at least as stable as the original method. Setting $\underline{\delta}_n = \hat{\underline{u}}_n - \underline{u}_n$, we have that

$$\underline{r}_0 = \underline{\delta}_0,$$

$$\underline{r}_n = \underline{\delta}_n - S\,\underline{\delta}_{n-1} - h\,\tilde{\underline{\delta}}_{n-1} \quad , \quad n = 1(1)N,$$

where $|\,\tilde{\underline{\delta}}_{n-1}\,| \leq L\,|\,\underline{\delta}_{n-1}\,|$. Thus

$$\sum_{j=0}^{n} S_0^{n-j}\,\underline{r}_j = S_0^n\,\underline{\delta}_0 + \sum_{j=1}^{n} S_0^{n-j}(\underline{\delta}_j - S\,\underline{\delta}_{j-1} - h\,\tilde{\underline{\delta}}_{j-1}),$$

and since $S_0^2 = S_0$ and $S_0\,S = S_0$,

$$\begin{aligned} |\,\sum_{j=0}^{n} S_0^{n-j}\,\underline{r}_j\,| &= |\,\underline{\delta}_n + (S_0 - S)\,\underline{\delta}_{n-1} - h(\sum_{j=1}^{n} S_0^{n-j}\,\tilde{\underline{\delta}}_{j-1})\,| \\ &\leq (1 + |\,S_0 - S\,| + t_n\,|\,S_0\,|\,L) \max_{0 \leq n \leq N} |\,\underline{\delta}_n\,|. \end{aligned}$$

Therefore by Theorem 3.6

$$\| \hat{\underline{U}}^{(0)} - \underline{U}^{(0)} \|_\infty \leq C_0 \| [S_0]^{-1} \underline{R} \|_\infty$$

$$\leq C_0 (1 + | S_0 - S | + | S_0 | L) \| \hat{\underline{U}} - \underline{U} \|_\infty,$$

which shows that (4.3) is at least as stable as the original method. Having established this result, let us show the necessity of the VSRC for optimal stability. Assume the original method is optimally stable. Then it must be at least as stable as (4.3), and so by Lemma 4.1 the original method must satisfy the VSRC. Next we show that the VSRC is sufficient for optimal stability. Assume the original method satisfies the VSRC. Let any other convergent method (Method 2) be given which is at least as stable as the original method. We must show that the original method is at least as stable as Method 2. From Lemma 4.1 it follows that Method 2 satisfies the VSRC and that $E_2 E = E_2$. Hence the null space of E is a subspace of the null space of E_2. Because both methods are convergent, rank E_2 = s = rank E; so the null spaces of E and E_2 are identical. From $E_2 E = E_2$ it follows that

$$E_2(E E_2 - I) = 0,$$

and since E and E_2 have identical null spaces,

$$E(E E_2 - I) = 0,$$

which simplifies to

$$E E_2 = E.$$

Therefore

$$\max_{0 \leq n \leq N} \left| \sum_{j=0}^{n} E^{n-j} \underline{r}_j \right| = \max_{0 \leq n \leq N} \left| \sum_{j=0}^{n} E_2^{n-j} \underline{r}_j + (E - E_2) \sum_{j=0}^{n-1} E_2^{n-j} \underline{r}_j \right|$$

$$\leq (1 + | E_2 - E |) \max_{0 \leq n \leq N} \left| \sum_{j=0}^{n} E_2^{n-j} \underline{r}_j \right|,$$

and so

$$\begin{aligned} \| \hat{\underline{U}} - \underline{U} \|_\infty &\leq C \| [E]^{-1} \underline{R} \|_\infty \\ &\leq C (1 + | E_2 - E |) \| [E_2]^{-1} \underline{R} \| \\ &\leq C (1 + | E_2 - E |) c_2^{-1} \| \hat{\underline{U}}^{(2)} - \underline{U}^{(2)} \|_\infty, \end{aligned}$$

showing that the original method is at least as stable as Method 2. Q.E.D.

5. Order of convergence. In this section it is shown that the order of convergence is exactly equal to the order of quasi-consistency if the SRC is satisfied. Then examples are given of methods for which the order of quasi-consistency is greater than the order of consistency.

DEFINITION 5.1. A method is said to be convergent of order p for the problem (2.1) if

$$|| \underline{U} - \underline{Y} ||_\infty = O(h^p).$$

If a method is convergent of order p for all infinitely smooth problems (2.1), it is simply said to be convergent of order p.

DEFINITION 5.2. A method satisfying the RC is said to be quasi-consistent of order p with the problem (2.1) if

$$\max_{0 \le n \le N} | \underline{d}_n | = O(h^p)$$

and

$$\max_{0 \le n \le N} | E(\underline{d}_0 + \underline{d}_1 + \ldots + \underline{d}_n) | = O(h^p)$$

If this is true for all infinitely smooth problems (2.1), then the method is simply said to be quasi-consistent of order p.

For a method to be quasi-consistent of order p with a problem it is sufficient that

$$\text{(i)} \quad \underline{d}_n = O(h^p) \quad , \quad n = O(1)N \quad ,$$

$$\text{(ii)} \quad E\,\underline{d}_n = O(h^{p+1}) \quad , \quad n = 1(1)N.$$

THEOREM 5.3. *A method satisfying the SRC is convergent of order* p *for the problem* (2.1) *if and only if it is quasi-consistent of order* p *with* (2.1).

The proof is omitted since it is similar to that of Theorem 3.8.

Example 1. If the Adams-Bashforth-Moulton PEC method is written in the first way shown in Example 2 of §2, we have

$$E = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad \underline{d}_n = h^3 \begin{bmatrix} 0 \\ -5y_n^{(3)}/12 \\ 0 \end{bmatrix} + O(h^4),$$

and $E\,\underline{d}_n = O(h^4)$, showing that the method is quasi-consistent of order three. Written in the second form, the method is also consistent of order three.

Example 2. Butcher's three-stage procedure of order four has

$$E = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad \underline{d}_n = h^4 \begin{bmatrix} -\frac{5}{96} f_y(t_n, y_n)\, y_n^{(3)} \\ 0 \end{bmatrix} + O(h^5)$$

and $E\,\underline{d}_n = O(h^5)$. There seems to be no form of this method for which the order of consistency is four.

Example 3. The modified linear multistep method has

$$E = \begin{bmatrix} 1/2 & 5/6 & 1/2 & 1/6 \\ 0 & 0 & 0 & 0 \\ 1/2 & 5/6 & 1/2 & 1/6 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \underline{d}_n = h^4 \begin{bmatrix} y_n^{(4)}/48 \\ 0 \\ -y_n^{(4)}/48 \\ 0 \end{bmatrix} + O(h^5)$$

and $E\,\underline{d}_n = O(h^5)$.

Example 4. The optimally stable two-cyclic method which alternates Milne's formula with the third order Adams-Moulton formula (see Stetter [7, p. 217]) is convergent of order four yet only consistent of order three.

6. Asymptotic behavior of the global discretization error. The principal term in the asymptotic expansion of the global error is given in Theorem 6.1 for methods satisfying the SRC. There is some simplification if, in fact, the VSRC is satisfied. This is discussed after the proof of the theorem.

THEOREM 6.1. *Assume* $\underline{\psi}_{\underline{u}}(t, \underline{u}; h)$ *is uniformly Lipschitz continuous in* $\underline{u}$ *and continuous in* t *and* h *for* $0 \le t \le 1$, $\underline{u} \in R^k$, *and* $0 \le h \le h_f$. *Assume*

$$\underline{d}_0 = h^p \underline{\gamma}_p + o(h^p),$$

$$\underline{d}_n = h^p \underline{\phi}_p(t_n) + h^{p+1} \underline{\phi}_{p+1}(t_n) + o(h^{p+1}), \qquad n = 1(1)N,$$

where $\underline{\phi}_p(t)$ *and* $\underline{\phi}_{p+1}(t)$ *are continuous and* $E \underline{\phi}_p(t) \equiv 0$. *Then*

$$\underline{u}_n = \underline{y}_n + h^p(\underline{\varepsilon}(t_n) + (I - S + E)^{-1} \underline{\phi}_p(t_n) + (S - E)^n \underline{c}) + o(h^p)$$

uniformly for $0 \le n \le N$ *where*

$$\underline{c} = (I - E) \underline{\gamma}_p - (I - S + E)^{-1} \underline{\phi}_p(0)$$

and the function $\underline{\varepsilon}(t)$ *satisfies*

$$\underline{\varepsilon}(0) = E \underline{\gamma}_p$$

$$\underline{\varepsilon}' = E\, G(t) (\underline{\varepsilon} + (I - S + E)^{-1} \underline{\phi}_p(t)) + E \underline{\phi}_{p+1}(t)$$

with

$$G(t) = \underline{\psi}_{\underline{u}}(t, \Lambda y(t); 0).$$

Proof. Let

$$\hat{\underline{u}}_n = \underline{y}_n + h^p(\underline{\varepsilon}(t_n) + (I - S + E)^{-1} \underline{\phi}_p(t_n) + (S - E)^n \underline{c}).$$

Then we must have

$$\underline{r}_0 = \hat{\underline{u}}_0 - \underline{\sigma}(\overline{y}_0; h),$$

$$\underline{r}_n = \hat{\underline{u}}_n - S \hat{\underline{u}}_{n-1} - h \underline{\psi}(t_{n-1}, \hat{\underline{u}}_{n-1}; h), \qquad n = 1(1)N.$$

It is easily verified that $\underline{c}$ has been chosen such that $\underline{r}_0 = o(h^p)$. By the

mean value theorem and the Lipschitz continuity of $\underline{\psi}_{\underline{u}}$

$$\underline{r}_n = \hat{\underline{u}}_n - S\,\hat{\underline{u}}_{n-1} - h\,\underline{\psi}(t_{n-1}, \underline{y}_{n-1}; h)$$
$$-h\,\underline{\psi}_{\underline{u}}(t_{n-1}, \underline{y}_{n-1}; h)\,(\hat{\underline{u}}_{n-1} - \underline{y}_{n-1}) + O(h^{2p+1})$$

for $n \geq 1$. It follows that

$$\underline{r}_n = -\underline{d}_n + (\hat{\underline{u}}_n - \underline{y}_n) - (S + h\,G(t_{n-1}))\,(\hat{\underline{u}}_{n-1} - \underline{y}_{n-1}) + o(h^{p+1}). \tag{6.1}$$

Appropriate substitutions yield

$$\underline{r}_n = h^p\{-\underline{\phi}_p(t_n) + \underline{\varepsilon}(t_n) + (I - S + E)^{-1}\,\underline{\phi}_p(t_n) + (S - E)^n\,\underline{c}$$
$$- S\,\underline{\varepsilon}(t_{n-1}) - S(I - S + E)^{-1}\,\underline{\phi}_p(t_{n-1}) - S(S - E)^{n-1}\,\underline{c}\} + O(h^{p+1}). \tag{6.2}$$

Because $E\,\underline{\phi}_p(t) = \underline{0}$ and $(I - S)\,(I - S + E)^{-1} = I - E$,

$$-\underline{\phi}_p(t) + (I - S + E)^{-1}\,\underline{\phi}_p(t) = S(I - S + E)^{-1}\,\underline{\phi}_p(t).$$

Also $S\,\underline{\varepsilon}(t) = \underline{\varepsilon}(t)$ and $E\,\underline{c} = \underline{0}$, and so (6.2) becomes

$$\underline{r}_n = h^p(\underline{\varepsilon}(t_n) - \underline{\varepsilon}(t_{n-1})) + h^p\,S(I - S + E)^{-1}\,(\underline{\phi}_p(t_n) - \underline{\phi}_p(t_{n-1})) + O(h^{p+1}).$$

Thus by continuity $\underline{r}_n = o(h^p)$. From (6.1)

$$E\,\underline{r}_n = h^{p+1}\,E\{-\underline{\phi}_{p+1}(t_n) + h^{-1}\,\underline{\varepsilon}(t_n) - h^{-1}\,\underline{\varepsilon}(t_{n-1})$$
$$- G(t_{n-1})\,[\underline{\varepsilon}(t_{n-1}) + (I - S + E)^{-1}\,\underline{\phi}_p(t_{n-1}) + (S - E)^{n-1}\,\underline{c}]\} + o(h^{p+1}).$$

Making use of the differential equation for $\underline{\varepsilon}(t)$ gives

$$E\,\underline{r}_n = h^{p+1}\,E\{h^{-1}\,\underline{\varepsilon}(t_n) - h^{-1}\,\underline{\varepsilon}(t_{n-1}) - \underline{\varepsilon}'(t_{n-1}) - G(t_{n-1})\,(S - E)^{n-1}\,\underline{c}\} + o(h^{p+1})$$
$$= -h^{p+1}\,E\,G(t_{n-1})\,(S - E)^{n-1}\,\underline{c} + o(h^{p+1}).$$

Therefore since

$$\sum_{n=1}^{\infty} |\,(S - E)^{n-1}\,| < \infty\ ,$$

we must have

$$|\,\sum_{j=0}^{n-1} E\,\underline{r}_j + \underline{r}_n\,| = o(h^p),$$

which by Theorem 3.6 is equivalent to $|| \hat{\underline{U}} - \underline{U} ||_\infty = o(h^p)$. Q.E.D.

In the case of a convergent method satisfying the VSRC we have that $E = \Lambda M^T$. Differentiating (3.7) with respect to y yields

$$M^T \underline{\psi}_u(t, \Lambda y; 0) \Lambda = f_y(t, y),$$

and hence

$$\begin{aligned} E\, G(t)\, \underline{\varepsilon}(t) &= E\, G(t)\, E\, \underline{\varepsilon}(t) \\ &= \Lambda\, f_y(t, y(t))\, M^T \underline{\varepsilon}(t). \end{aligned}$$

If we define $\varepsilon(t) = M^T \underline{\varepsilon}(t)$, then the magnified error function is

$$\Lambda\, \varepsilon(t) + (I - S + E)^{-1} \underline{\phi}_p(t)$$

where

$$\varepsilon(0) = M^T \underline{\gamma}_p$$

and

$$\varepsilon' = f_y(t, y(t))\, \varepsilon + M^T[G(t)\, (I - S + E)^{-1} \underline{\phi}_p(t) + \underline{\phi}_{p+1}(t)].$$

Note that the stability of the differential equation for $\varepsilon(t)$ depends only on the problem and not on the method.

COROLLARY 6.2. *Let the hypotheses of Theorem* 6.1 *be satisfied, and let* $0 < \theta \leq 1$ *be given. Then*

$$\underline{u}_n = \underline{y}_n + h^p(\underline{\varepsilon}(t_n) + (I - S + E)^{-1} \underline{\phi}_p(t_n)) + o(h^p)$$

uniformly for $\theta N \leq n \leq N$.

Proof. For $\theta N \leq n \leq N$ we have

$$\begin{aligned} | (S - E)^n \underline{c} | &\leq B\rho^n && \text{for some } 0 < \rho < 1 \\ &\leq B\rho^{\theta N} \\ &< B\rho^{-\theta q}\, \rho^{\theta/h}, \end{aligned}$$

which goes to zero faster than any power of h. Q.E.D.

Remark. Gragg (see Stetter [7, p. 234]) has shown that for p-th order convergent linear multistep methods satisfying the SRC the error posesses an

asymptotic expansion of the form

$$u_n = y(t_n) + \sum_{j=p}^{J} h^j \varepsilon_j(t_n) + O(h^{J+1})$$

uniformly for $\theta N \leq n \leq N$ as long as $f(t, y)$ posesses J-th partial derivatives which are uniformly Lipschitz continuous. It is reasonable to expect this result to be valid for all fixed-stepsize methods satisfying the SRC as long as the increment function is smooth enough.

Example 1. For the straight multistep method

$$\underline{d}_n = h^{p+1} \underline{e}_1 \phi_{p+1}(t_n) + O(h^{p+2}).$$

By the remarks following Theorem 6.1 the magnified error function is

$$\underline{\varepsilon}(t) = \underline{e}\, \varepsilon(t)$$

where $\varepsilon(t)$ solves the initial value problem

$$\varepsilon(0) = \bar{\rho}(1)^{-1} [\varepsilon_{q-1} + (1 + \alpha_1) \varepsilon_{q-2} + \ldots + (1 + \alpha_1 + \ldots + \alpha_{q-1}) \varepsilon_0],$$

$$\varepsilon' = f_y(t, y(t))\, \varepsilon + \bar{\rho}(1)^{-1} \phi_{p+1}(t)$$

assuming

$$u_i = y(t_i) + h^p \varepsilon_i + O(h^{p+1}), \qquad i = 0(1)q - 1.$$

Example 2. For Milne's method

$$d_n = h^5 \begin{bmatrix} 1 \\ 1 \end{bmatrix} \frac{y_n^{(5)}}{2880} + O(h^6),$$

$$\underline{\psi}(t, \underline{u}; 0) = \begin{bmatrix} 1/3 & 2/3 \\ 2/3 & 1/3 \end{bmatrix} \begin{bmatrix} f(t, u^1) \\ f(t, u^2) \end{bmatrix},$$

$$\underline{\psi}_{\underline{u}}(t, \underline{u}; 0) = \begin{bmatrix} 1/3 & 2/3 \\ 2/3 & 1/3 \end{bmatrix} \begin{bmatrix} f_y(t, u^1) & 0 \\ 0 & f_y(t, u^2) \end{bmatrix},$$

$$G(t) = f_y(t, y(t)) \begin{bmatrix} 1/3 & 2/3 \\ 2/3 & 1/3 \end{bmatrix}.$$

The magnified error function is $\underline{\varepsilon}(t)$ where

$$\underline{\varepsilon}' = f_y(t, y(t)) \begin{bmatrix} 1/3 & 2/3 \\ 2/3 & 1/3 \end{bmatrix} \underline{\varepsilon} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} \frac{y^{(5)}(t)}{2880} .$$

This system can be decoupled:

$$\underline{\varepsilon} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \varepsilon^1 \\ \varepsilon^2 \end{bmatrix}$$

where

$$(\varepsilon^1)' = f_y(t, y(t))\, \varepsilon^1 + \frac{y^{(5)}(t)}{2880} ,$$

$$(\varepsilon^2)' = -\frac{1}{3} f_y(t, y(t))\, \varepsilon^2 .$$

Example 3. For the modified linear multistep method equivalent to the fourth order Adams-Moulton method

$$E = \begin{bmatrix} 1/2 & 5/6 & 1/2 & 1/6 \\ 0 & 0 & 0 & 0 \\ 1/2 & 5/6 & 1/2 & 1/6 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad (I - S + E)^{-1} = \begin{bmatrix} 1 & 1/8 & 0 & -1/24 \\ 0 & 1 & 0 & 0 \\ 0 & -11/24 & 1 & 1/24 \\ 0 & 1 & 0 & 1 \end{bmatrix},$$

$$\underline{d}_n = h^4 \begin{bmatrix} 1/48 \\ 0 \\ -1/48 \\ 0 \end{bmatrix} y_n^{(4)} + h^5 \begin{bmatrix} -1/80 \\ 0 \\ 17/720 \\ 0 \end{bmatrix} y_n^{(5)} + h^5 \begin{bmatrix} 1/128 \\ 1/48 \\ -1/1152 \\ 0 \end{bmatrix} f_y(t_n, y_n)\, y_n^{(4)} + O(h^6) ,$$

$$\underline{\psi}(t, \underline{u}; 0) = \begin{bmatrix} 3/8 \\ 1 \\ -1/24 \\ 0 \end{bmatrix} f(t, \tfrac{1}{2} u^1 + u^2 + \tfrac{1}{2} u^3 + \tfrac{1}{8} u^4),$$

$$\underline{\psi}_{\underline{u}}(t, \underline{u}; 0) = \begin{bmatrix} 3/8 \\ 1 \\ -1/24 \\ 0 \end{bmatrix} f_y(t, \tfrac{1}{2}u^1 + u^2 + \tfrac{1}{2}u^3 + \tfrac{1}{8}u^4) \begin{bmatrix} \frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{8} \end{bmatrix},$$

$$G(t) = \begin{bmatrix} 3/8 \\ 1 \\ -1/24 \\ 0 \end{bmatrix} f_y(t, y(t)) \begin{bmatrix} \frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{8} \end{bmatrix}.$$

The magnified error function is

$$\begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \varepsilon(t) + \begin{bmatrix} 1/48 \\ 0 \\ -1/48 \\ 0 \end{bmatrix} y^{(4)}(t)$$

where

$$\varepsilon(0) = \tfrac{1}{2}\gamma_4^1 + \tfrac{5}{6}\gamma_4^2 + \tfrac{1}{2}\gamma_4^3 + \tfrac{1}{6}\gamma_4^4$$

and

$$\varepsilon' = f_y(t, y(t))\ \varepsilon + \frac{1}{180} y^{(5)}(t) + \frac{1}{48} f_y(t, y(t))\ y^{(4)}(t)$$

This can be rewritten

$$(\varepsilon + \tfrac{1}{48} y^{(4)}(t))' = f_y(t, y(t))\ (\varepsilon + \tfrac{1}{48} y^{(4)}(t)) + \frac{19}{720} y^{(5)}(t)\ ;$$

hence $\varepsilon(t) + \frac{1}{48} y^{(4)}(t)$ satisfies the differential equation for the magnified error function of the fourth order Adams-Moulton method.

REFERENCES

[1] B. Chartres and R. Stepleman, *A general theory of convergence for numerical methods*, SIAM J. Numer. Anal., 9 (1972), pp. 476-492.

[2] C. W. Gear, *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, N.J., 1971.

[3] P. Henrici, *Discrete Variable Methods for Ordinary Differential Equations*, John Wiley, New York, 1962.

[4] P. Lancaster, *Theory of Matrices*, Academic Press, New York, 1969.

[5] A. Nordsieck, *On the numerical integration of ordinary differential equations*, Math. Comp., 16 (1962), pp. 22-49.

[6] M. Spijker, *On the structure of error estimates for finite difference methods*, Numer. Math., 18 (1971), pp. 73-100.

[7] H. J. Stetter, *Analysis of Discretization Methods for Ordinary Differential Equations*, Springer-Verlag, New York, 1973.

Form AEC–427
(6/68)
AECM 3201

U. S. ATOMIC ENERGY COMMISSION
UNIVERSITY–TYPE CONTRACTOR'S RECOMMENDATION FOR DISPOSITION OF SCIENTIFIC AND TECHNICAL DOCUMENT

(See Instructions on Reverse Side)

1. AEC REPORT NO.	2. TITLE
COO-2383-0017	ANALYSIS OF FIXED-STEPSIZE METHODS

3. TYPE OF DOCUMENT (Check one):

- [X] a. Scientific and technical report
- [ ] b. Conference paper not to be published in a journal:
  - Title of conference _____
  - Date of conference _____
  - Exact location of conference _____
  - Sponsoring organization _____
- [ ] c. Other (Specify) _____

4. RECOMMENDED ANNOUNCEMENT AND DISTRIBUTION (Check one):

- [X] a. AEC's normal announcement and distribution procedures may be followed.
- [ ] b. Make available only within AEC and to AEC contractors and other U.S. Government agencies and their contractors.
- [ ] c. Make no announcement or distrubution.

5. REASON FOR RECOMMENDED RESTRICTIONS:

6. SUBMITTED BY: NAME AND POSITION (Please print or type)

C. W. Gear
Professor and Principal Investigator

Organization

Signature	Date
Charles W. Gear	February, 1975

FOR AEC USE ONLY

7. AEC CONTRACT ADMINISTRATOR'S COMMENTS, IF ANY, ON ABOVE ANNOUNCEMENT AND DISTRIBUTION RECOMMENDATION:

8. PATENT CLEARANCE:

- [ ] a. AEC patent clearance has been granted by responsible AEC patent group.
- [ ] b. Report has been sent to responsible AEC patent group for clearance.
- [ ] c. Patent clearance not required.

BIBLIOGRAPHIC DATA SHEET	1. Report No. UIUCDCS-R-75-702	2.	3. Recipient's Accession No.

4. Title and Subtitle

ANALYSIS OF FIXED-STEPSIZE METHODS

5. Report Date: February 1975

6.

7. Author(s)
Robert Skeel

8. Performing Organization Rept. No. UIUCDCS-R-75-702

9. Performing Organization Name and Address
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801

10. Project/Task/Work Unit No.

11. Contract/Grant No.
US AEC AT (11-1) 2383

12. Sponsoring Organization Name and Address
US Atomic Energy Commission
Chicago Operations Office
9800 South Cass Avenue
Argonne, Illinois 60439

13. Type of Report & Period Covered

14.

15. Supplementary Notes

16. Abstracts

The unified theory developed by Henrici [3] for one-step methods is extended to the more general case where the order of the system of difference equations can exceed the order of the system of ordinary differential equations. The analysis is applicable to every fixed-stepsize fixed-formula method known to the author. For many of these methods the concept of consistency is inadequate. A more appropriate concept, termed quasi-consistency, is introduced.

17. Key Words and Document Analysis. 17a. Descriptors

7b. Identifiers/Open-Ended Terms

7c. COSATI Field/Group

18. Availability Statement	19. Security Class (This Report)	21. No. of Pages
Unlimited	UNCLASSIFIED	
	20. Security Class (This Page) UNCLASSIFIED	22. Price